AD_____
)

Award Number:
W81XWH-07-1-0619

TITLE:
Genetic and Environmental Pathways in Type 1 Diabetes Complications

PRINCIPAL INVESTIGATOR:
Massimo Trucco, M.D.

CONTRACTING ORGANIZATION:
Children's Hospital of Pittsburgh
Pittsburgh, PA 15224

REPORT DATE:
September 2010

TYPE OF REPORT:
Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: (Check one)

X Approved for public release; distribution unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
09-26-2010	Annual	27 August 2009 – 26 August 2010

4. TITLE AND SUBTITLE
Genetic and Environmental Pathways in Type 1 Diabetes Complications

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-07-1-0619

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Massimo Trucco, M.D.

Email: mnt@pitt.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Children's Hospital of Pittsburgh of
UPMC Health System
One Children's Hospital Drive
4401 Penn Avenue
Pittsburgh, PA 15224

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research
and Materiel Command
Fort Detrick, MD 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Genetic factors contribute to risk for developing nephropathy in patients with Type 1 Diabetes (T1D). Cigarette smoking is deleterious to kidney function and is a risk factor for Diabetic-Nephropathy (DN) as well as End Stage Renal Disease (ESRD) in patients with T1D. The proposed study investigates how environmental exposure(s) (e.g., smoking) and genetic variants interact to amplify risk for T1DN and substantially increase incidence of ESRD. The specific aims are: 1) Identify genetic variants conferring risk to T1DN by performing a staged follow-up of our initial Genome-Wide Association Scan (GWAS) results; 2) Ensure that SNPs identified by Aim 1 affect risk of T1DN, as opposed to risk for T1D; 3) Identify genetic variants that interact with smoking status in conferring risk for T1DN; 4) Confirm results obtained during Aims 1-3 using an independent cohort of case and control participants. The relevance of the study to public health is that 16 million people in the US have diabetes with 800,000 new cases diagnosed each year. Diabetic complications threatening vision, kidney, and nerve function affect most diabetic patients. Improved prediction of risk for developing diabetes and diabetic complications among active duty members of the military, their families and retired military personnel will potentially allow focused preventative treatment of at-risk individuals, providing significant healthcare savings and improved patient well being.

15. SUBJECT TERMS
End Stage Renal Disease; Genetic Association; Genome Scanning; Nephropathy; Type 1 Diabetes

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
U	U	U	UU	22	**19b. TELEPHONE NUMBER** *(include area code)*

Children's Hospital of Pittsburgh
W81XWH-07-1-0619
Annual Report (08/27/2009 – 08/26/2010)
Table of Contents

INTRODUCTION:

Type 1 Diabetes (T1D) is associated with increased risk of T1D-Nephropathy (T1DN) and is usually accompanied by other diabetic-related complications such as retinopathy, neuropathy, blood pressure elevation, and high risk of cardiovascular morbidity and mortality. Sixteen million people in the US have diabetes with 800,000 new cases diagnosed each year. Diabetic complications affect most diabetic patients. Diabetes occurs in men, women, children and the elderly. African, Hispanic, Native and Asian Americans are particularly susceptible to its most severe complications. An estimated 20% to 40% of T1D patients will develop diabetic nephropathy, clinically first evidenced by microalbuminuria, during their lifetime. If untreated nearly all T1D patients experiencing microalbuminuria will progress to overt nephropathy, evidenced by macroalbuminuria, and culminating in T1D-End Stage Renal Disease (T1D-ESRD). Improved prediction of risk for developing diabetes and diabetic complications among active duty members of the military, their families and retired military personnel will potentially allow focused preventative treatment of at-risk individuals, providing significant healthcare savings and improved patient well being.

BODY:

Our first quarterly scientific progress report for the third year of our project (08/27/09 – 11/30/09) detailed the following steps forward in reaching the aims of our study.

In previous Quarterly Reports we addressed an expanded focus of the project to include proteomics research. This aspect of the expanded project is designed to quantify proteins in blood that can be linked to onset of Type 1 Diabetes (T1D). The goal is to identify diagnostic biomarkers that can be used alone or along with DNA genotyping to gauge risk for developing T1D and diabetes complications.

Rationale: The Diabetes Prevention Trial of Type 1 (DPT-1) observed that individuals with 2 or more autoantibodies exhibited an overall 68% 5 years risk while those with 3 autoantibodies approached a near absolute risk for developing T1D (Verge et al., 1996). Results from subsequent studies have shown that multiple autoantibodies confer cumulative risk ranging from 75% to nearly 90% during 5-10 year prospective follow up of first degree relatives (Pietropaolo and Becker, 2001). While screening for the presence of 3 autoantibodies has high predictive value a substantial minority (i.e., roughly 20% of first degree relatives who convert to T1D) exhibit fewer than 3 autoantibodies (Pihoker et al., 2005). Moreover, autoantibodies may also appear successively during disease progression and reflect ongoing autoimmune destruction of insulin producing pancreatic beta cells.

Autoantibodies provide a good indication of risk for developing T1D. However, the test does not allow staging of subjects for when in the next decade they will exhibit diabetes, it does not correlate with severity of the initial symptoms of diabetes, nor does it correlate with the patient's ability to control their diabetes. The presence of multiple autoantibodies is an indication that autoimmune destruction of insulin producing cells is already occurring (i.e., autoimmune disease is already present). Studies of subjects who are positive for multiple autoantibodies but remain diabetes free (i.e., false positive) indicate that for 2 autoantibodies roughly half the subjects will remain free of diabetes and for 3 autoantibodies 10% to 25% will remain diabetes free. Moreover, screening for autoantibodies has a significant false negative rate. It has been estimated by different studies that 7% to 19% of new onset T1D patients are autoantibody negative (Wang et al., 2007). A critical goal for prevention is to suppress destruction of beta cells as early as possible. For these reasons discovery of new biomarkers that can be detected prior to the appearance of autoantibodies (or that provide a chronologically accurate prediction of when T1D will develop) would be of great value.

T1D is among the most common chronic diseases of childhood with prevalence and incidence estimated in children less than 20 years of age at 192/100,000 and 12.2/100,000, respectively (Jacobson et al., 1997). The overall prevalence of the disease among siblings of T1D probands is increased by 30-fold while among HLA identical siblings prevalence increases by more than 80-fold. Cohorts used in the T1D trials are typically recruited from among siblings of T1D probands and exclusion/inclusion criteria make use of whether 2 or more autoantibodies are present along with the subject's HLA genotype. Because the presence of multiple autoantibodies is used as primary criteria for inclusion these patients are at a stage where they already present autoimmune disease. This imposes increased difficulty on development of successful prevention strategies.

Of course, only 15% of new T1D cases occur among first-degree relatives. In other words, for 85% of those who will present T1D there is currently no strategy for predicting risk.

A test that would enable improved risk estimation would find immediate use in prediction and prevention trials. An issue to consider is that peak incidence of disease occurs at 10 to 14 years of age and therefore the most important trials will involve recruitment of children. The false positive rate associated with the autoantibody test makes it an unsatisfactory tool to use when screening young subjects for inclusion. This would be especially true in the event of a trial designed to test a preventative treatment.

What population would ideally be used in a large clinical trial setting? In the event that the project identifies a set of strongly predictive biomarkers an appropriate next step would be to approach TrialNet (see http://www.diabetestrialnet.org). The TrialNet organization is a multi center study with the goal of identifying subjects for T1D prevention and intervention trials. Children's Hospital of Pittsburgh (CHP) is already acting as a clinical center for the TrialNet natural history study (Mahon et al., 2009). TrialNet has currently assembled a large cohort of family members (N=12,636) of T1D index cases of which greater than 300 are being monitored in a prospective study for onset of T1D. Dr. Trucco is a member of the TrialNet Steering Committee and is also the Chair of the TrialNet Scientific Review Committee.

Specific Aims: The specific aim for the expanded project is to investigate proteins and analytes present in blood in order to determine if changes in abundance as well as changes in post-translational modifications are associated with increased risk for developing T1D. In a pilot experiment we will use serum collected from a cohort of T1D cases and non-T1D healthy controls to validate that the proteomics approach identifies changes in protein or analyte abundance associated with disease. This will be followed by an analysis of serum collected from subjects who converted to T1D. We will choose a set of serum samples collected at multiple time points from T1D cases. Blinded samples will be tested on aptamer arrays and quality control analysis of the resulting data. Following quality control analysis the samples will be unblinded. Data will be analyzed with the goal of identifying the optimal set of serum biomarkers exhibiting sensitivity to T1D risk.

In latter stages of the project we will select a second set of serum samples from T1D cases that can be used to validate markers discovered during Aim 1 of the project. Blinded samples will be tested and will be unblinded following quality control analysis. In this example only the biomarkers identified initially as being sensitive for T1D risk need to be analyzed. The final stage will be to analyze samples collected from low-risk participants. We will select a set of serum samples from low-risk individuals recruited in the CHP longitudinal cohort. These samples have been collected at multiple time points from participants who have remained non-T1D at the study's endpoint. Blinded samples will be used and following quality control steps will be unblinded.

Table 1. T1D Case and non-T1D Healthy Control Samples.

	N	Material	Source
Case	30	Serum	AOB Data/Serum Bank
Control	30	Serum	AOB Data/Serum Bank

Progress Report: We are planning an experiment in which we will use serum collected from thirty subjects with Type 1 Diabetes (T1D) and from thirty healthy (i.e., non-diabetic) subjects recruited from Italy (Table 1). The serum samples were originally collected by the Non Insulin Requiring Autoimmune Diabetes (NIRAD) Study and are being stored in the AOB Data/Serum Bank. Our experimental design is to work with 50ul of serum from each de-identified subject. The material received will be used to examine the abundance of circulating proteins and analytes present in blood with the goal of identifying proteins that can be used to diagnosis risk of developing T1D. These samples are already in existence but are no longer needed by the original study.

12. Statement of Plans for the Upcoming Research Period

Goal 1. Prepare serum samples for proteomic analysis on aptamer arrays.

Goal 2. Initiate proteomic analysis of serum collected from N=30 T1D cases and N=30 non-T1D controls.

Literature Cited:

Jacobson DL, Gange SJ, Rose NR, Graham NM. (1997) Epidemiology and estimated population burden of selected autoimmune diseases in the United States. Clin Immunol Immunopathol 84:223-243.

Mahon JL, Sosenko JM, Rafkin-Mervis L, Krause-Steinrauf H, Lachin JM, Thompson C, Bingley PJ, Bonifacio E, Palmer JP, Eisenbarth GS, Wolfsdorf J, Skyler JS; TrialNet Natural History Committee; Type 1 Diabetes TrialNet Study Group. (2008) The TrialNet Natural History Study of the Development of Type 1 Diabetes: objectives, design, and initial results. Pediatr Diabetes 10:97-104.

Pietropaolo M, Becker DJ. (2001) Type 1 diabetes intervention trials. Pediatr Diabetes 2:2-11.

Pihoker C, Gilliam LK, Hampe CS, Lernmark A. (2005) Autoantibodies in diabetes. Diabetes 54(Suppl 2):S52-61.

Verge CF, Gianani R, Kawasaki E, Yu L, Pietropaolo M, Jackson RA, Chase HP, Eisenbarth GS. (1996) Prediction of type I diabetes in first-degree relatives using a combination of insulin, GAD, and ICA512bdc/IA-2 autoantibodies. Diabetes 45:926-933.

Wang J, Miao D, Babu S, Yu J, Barker J, Klingensmith G, Rewers M, Eisenbarth GS, Yu L. (2007) Prevalence of autoantibody-negative diabetes is not rare at all ages and increases with older age and obesity. J Clin Endocrinol Metab 92:88-92.

In our second quarterly scientific progress report (12/01/09 – 02/28/10), we presented the following data:

During the previous research quarter our efforts focused upon two projects not mentioned in our prior quarterly reports. They were to (1) finalize a manuscript for publication and (2) follow up on the results obtained during the current DOD funded project by beginning the application process for extramural funding to support additional studies into the genetics of inherited diabetes. This quarterly report will summarize our effort to accomplish these two goals.

Completion of Goal 1. Finalize a manuscript describing the development of statistical tools for identifying genomic variants (i.e., single nucleotide polymorphisms, SNPs) that affect an individual's susceptibility to disease. The manuscript has been submitted to the peer-reviewed journal ANNALS OF STATISTICS.

The manuscript describes our research into development of methods for combining genetic data garnered from family-based studies with those collect from unrelated subjects during case-control comparisons. These approaches represent the primary sampling techniques for studies of gene to phenotype association. Due to demographic, biological, and random forces, genetic variants differ in allele frequency in populations around the world, creating population structure or stratification reflected by ancestry. As a consequence, case-control studies are susceptible to spurious associations between genetic variants and disease status (Lander et al., 1994). As more data are collected the challenge of spurious associations due to population structure increases (Devlin and Roeder, 1999; Devlin et al., 2001; Devlin et al., 2004). The research problem addressed in our recently submitted publication is how to use both case-control and family-based data in a single test for association. We seek to develop an approach in which the test is more powerful and robust to population stratification than competing approaches (Nagelkerke et al., 2004; Epstein et al., 2005). Our approach, which is robust to differences in sampling distribution across studies, control Type I error while attaining good power. The method requires that sufficient genotyping is available on all samples to permit matching samples based on genetic ancestry.

As a first step we estimated the genetic background of unrelated individuals (cases, controls, and trio probands). We considered genotypic data from the International HapMap Project (30 CEU trios) and

from the POPRES database (Nelson et al., 2008). Trio probands are matched to one or more controls that are genetically similar (Figure 1) (Luca et al., 2008). The distance between individuals in this eigenspace is representative of their genetic differences. When data consist of family-based samples as trios of parents and their affected offspring, as well as additional controls, we will prefer matching one case to many controls. For trios pseudo-controls are automatically matched by ancestry with the corresponding proband, and will be contrasted to the case genotype. Additional information can be garnered by clustering trio probands with unrelated controls. In this way we identify additional controls matched by ancestry to the probands (Figure 1). The structure of the data is equivalent to a matched case-control sample and hence can be analyzed via conditional logistic regression.

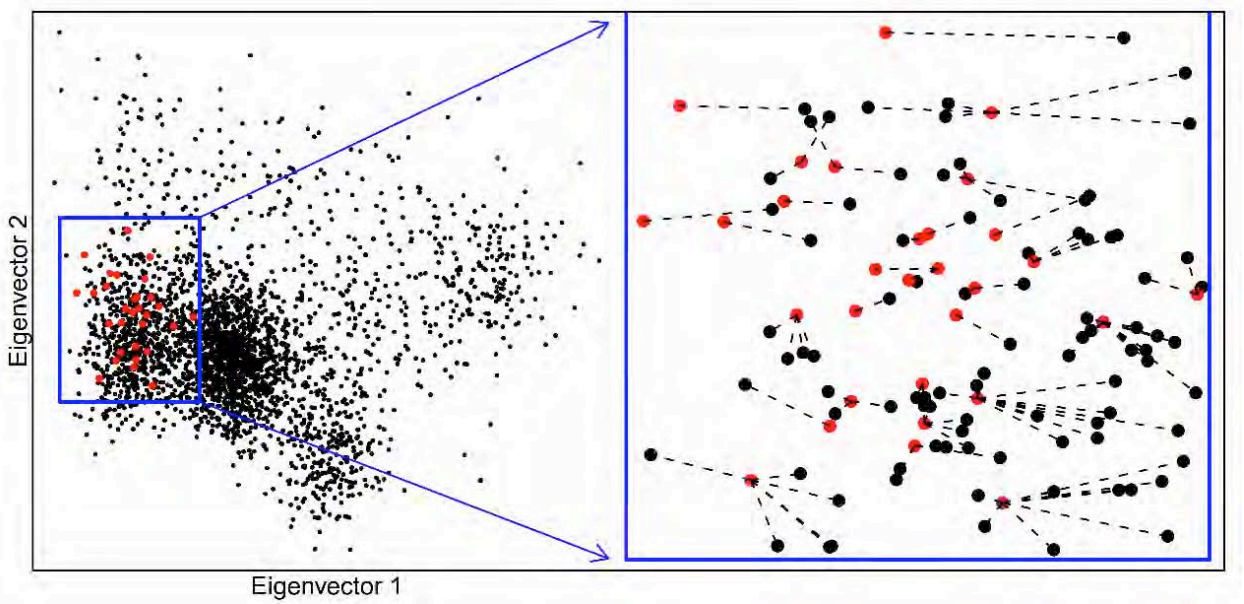


Figure 1. HapMap trios matched by ancestry to POPRES controls. The 30 off spring from HapMap, CEU samples, trios serve as cases and the 2,184 individuals of European ancestry from the POPRES data serve as controls. The plot displays the top two principal components of ancestry for cases (red) and controls (black). Based on the distribution of points in the eigenmap, many available controls would not be good matches to the HapMap trios. Only those delineated in blue are considered further. Each case is matched to one or more controls that are genetically similar based on the eigenvectors.

Some unrelated controls will not be similar enough to any probands to merit inclusion in the study. For example, the HapMap trios can only be successfully matched to a subset of the full European samples in POPRES (Figure 1). Likewise some unrelated cases might not be well matched by any unrelated controls in the study. Our approach provides features that facilitate the removal of individuals who cannot be successfully matched because their genetic ancestry is too remote, relative to the others in the samples (Luca et al., 2008; Lee et al., 2010). These individuals should be removed from further consideration in the association study. Once the strata are established, a natural next step is to compare the differences in genotypes between the cases and controls by using conditional logistic regression (data not shown).

Application to Type 1 Diabetes. In previous studies Type 1 Diabetes (T1D) has demonstrated a strong association with the HLA region of chromosome 6 (Davies et al., 1994). To illustrate our method we consider joint analysis of 19 T1D trios with just over 2,000 independent controls. All family and control samples are of European ancestry; for details about the data see Luca et al. (Luca et al., 2008). First, we estimated the ancestry of the controls and plotted them against their two most significant axes of genetic variation (Figure 2). We then projected the 19 trio probands onto the control's eigenmap. The full match algorithm identified 19 distinct strata, each including exactly one trio proband, and

between 19 and 359 controls. We call these unbalanced strata "all controls", to indicate that we matched the full sample controls.

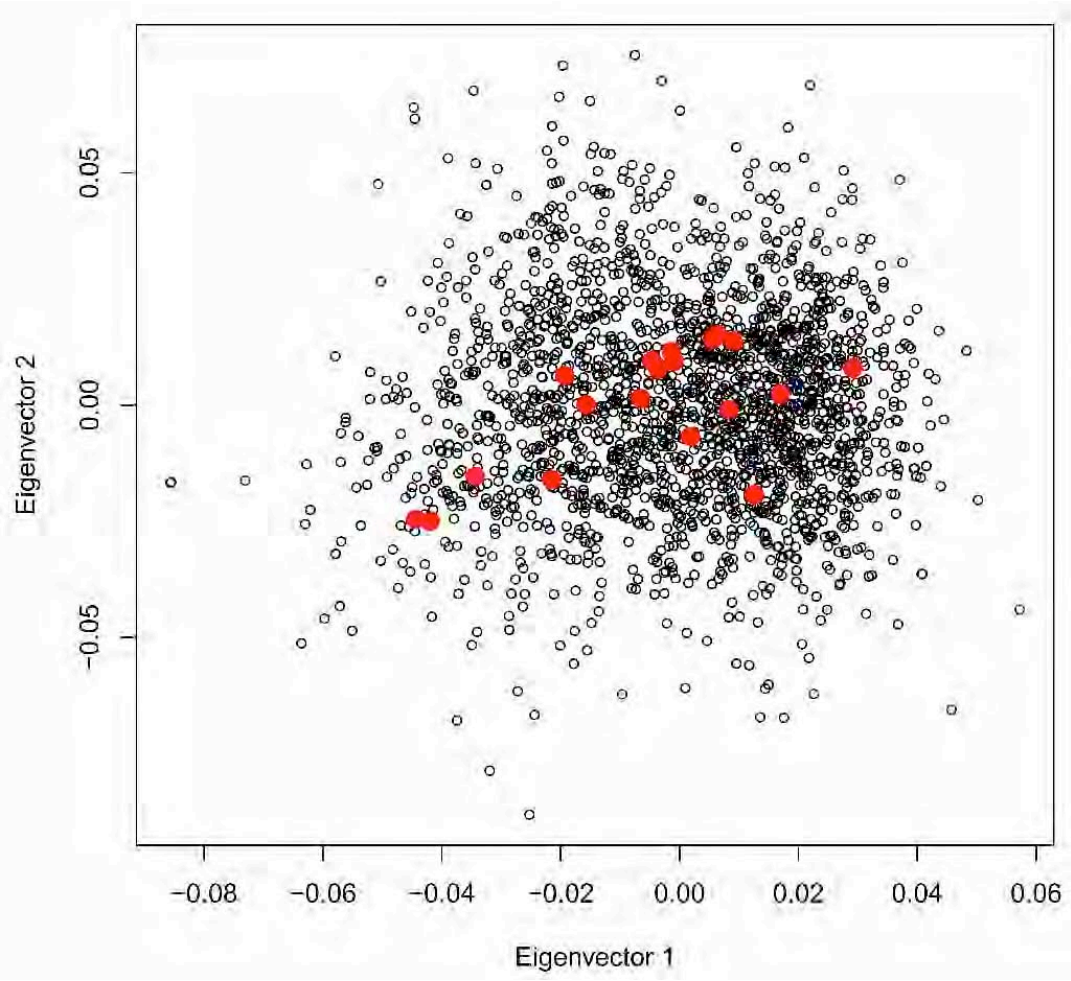


Figure 2. Eigenmap for Type 1 diabetes data. Probands (red) are plotted on the eigenmap determined by the controls (black).

For our analysis we also chose the closest controls to each case. For SNPs in the HLA region, we evaluated the success at detecting association with T1D. From our results it is apparent that as the number of matches increases the power to detect certain SNPs also increases (Figure 3). The best p-value is well over two orders of magnitude better when using all of the controls. The strongest signals occur at SNPs rs241427 and rs9273363 located near the confirmed T1D susceptibility locus HLA-DQB1 within the HLA class II region (Davies et al., 1994).

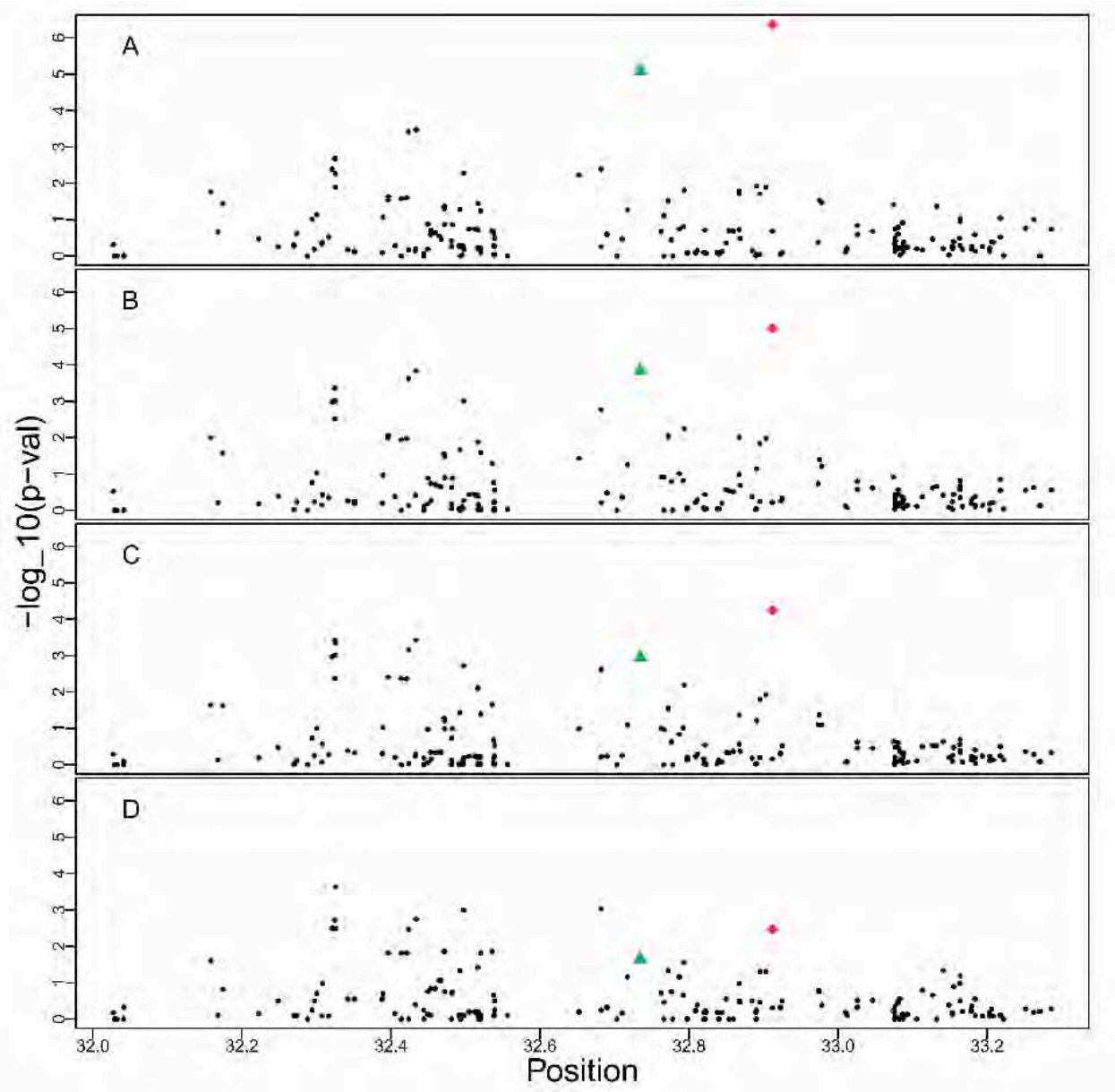


Figure 3. Association between HLA markers and Type 1 Diabetes. The -log10(p-values) are plotted versus individual SNPs in the HLA region of chromosome 6. (A) All controls matched; (B) 1:10matching: (C) 1:5 matching; (D) Trios only. The strongest association occurs for rs241427 (diamond) and next strongest for rs9273363 (triangle).

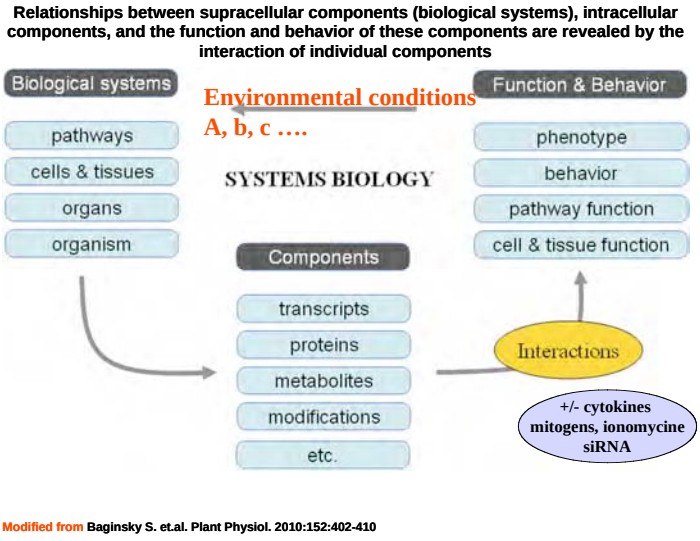
Completion of Goal 2. Initiate the application process for extramural funding.

We have identified the Human Frontiers Science Program (HFSP) along with the National Institutes of Health (NIH) as potential funding sources to support our future research into genetic influences affecting risk of developing diabetes. Our hypothesis is that highly penetrant gene variants can be identified through their effect on molecular networks. The goal of the new project will be to develop molecular and statistical tools to identify perturbations of expression in gene networks. Ultimately these tools will be useful for biologists searching genomes for rare, highly penetrant variants.

The project aims are as follows:
Project Aim 1: a) use gene and pathway ontology, and our experiments, to identify master and minor gene-expression regulators (leading indicator genes) in a model system (B-lymphocyte cell-lines) after stimulation and during dynamic growth; b) perturb select genes in co-regulated pathways by siRNA knockdown; c) measure RNA expression at multiple time points by massively parallel sequencing; d) build statistical models from resulting data.

Project Aim 2: test statistical models by additional experiments; refine and "robustify" models considering these results.

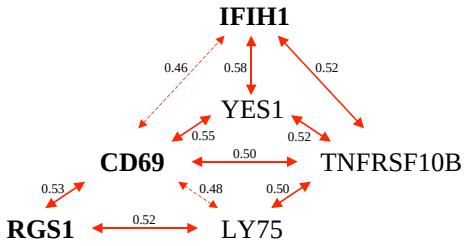
Project Aim 3: experiments paralleled by statistical models to determine if perturbation of expression can be detected from pooled samples, in which only a fraction of samples have perturbed gene expression. If such deconvolution is possible, it will reduce costs for experiments aimed at identifying rare variants affecting phenotype.



Project Summary. The project will analyze the biology of model cell lines (i.e., B-lymphocyte cell lines [BLCL]) strains. Data will be collected on siRNA knockdown of select genes and molecular networks. The BLCL can be used as surrogate cells for the study of antigen presenting cell function. Molecular analyses (i.e., immunogenetic and transcriptome data) will be conducted and used to synthesize a mathematical model of molecular and cellular data for correlations in molecular responses of BLCL to environmental conditions (i.e., activating stimuli). During the project model cell lines will be created to identify genes regulating molecular networks. Knockdown strains will be generated to study the response of BLCL to environmental conditions (e.g., cytokines, phorbol myristate acetate, ionomycine treatment). Our expertise in the creation and characterization of BLCL will be used to screen the already existing collection of cell lines (n=300) immortalized from a healthy ancestry matched cohort. Conditions evaluated will be EBV infection, growth rate, secretion of cytokines, expression of cell surface markers, and post-translational modifications of MAP kinase proteins.

The project will use next generation sequencing to evaluate changes in RNA abundance, splice variants, and allele specific expression. The resulting data will be integrated with time series data on cell phenotypic variation to support development of robust, predictive data models of the network phenotype associated with BLCL response under defined environmental conditions (e.g., various mitogen stimulation of BLCL, surrogate antigen presenting cells). The data will be combined with molecular aspects of the project and mathematical modeling to recognize molecular networks and network components (e.g., gene expression correlations) causal for network and cell response to environmental signals. Integration of biological data with bioinformatics and gene/pathway ontology and mathematical models will be used to develop comprehensive description of surrogate antigen presenting cell response to environmental stimuli. During the project we will compare data garnered from mitogen stimulated BLCL in the presence of specific MAP kinase inhibitors as well as siRNA knockdowns of network regulatory genes that will allow development of a data model in which changes to network phenotypes can be identified.

Interacting Networks and Common Variants



Red Arrows indicate positive Pearson correlation (average r=0.52)
Source: Nayak et al. (2009) Genome Research 19:1953-1962

Loci in LD with Common Variants Associated with Autoimmune Disease Phenotypes

Locus	Chromosome	Autoimmune Phenotype	Common Variant	Odds Ratio (95%CI)
RGS1	1q31.2	Celiac, T1D	rs2816316	0.89 (0.84-0.95)
CD69	12p13.31	T1D	rs4763879	1.09 (1.02-1.16)
IFIH1	2q24.2	SLE, T1D	rs1990760	0.86 (0.82-0.90)

Source: T1DBase (http://t1dbase.org)

Mathematical work will be an integral part of the project. We will build on the literature of Graphical Gaussian models and dynamic Bayesian networks to develop and use statistical methods to estimate co-expression networks for longitudinal data. The models aim to detect conditionally dependent genes from the experimental data and thus determine the causal relationships within networks. From these results, we will develop methods for modeling the effects of perturbations such as knockdowns on the networks. Based on the partial correlation structure before and after perturbation, we can construct hypothesis tests for which genes are regulating the network. These results will determine our power to detect perturbations when the signal is less refined. Pooled data, for example, will yield a muted signal due to the convex combination of network signals. Analysis of data, both simulated and actual, will determine the feasibility of pooling as a strategy.

Project Outcomes. Identify rare variants affecting phenotypic variability - identify environmental sensitive variants. Integrate levels of analysis ranging from mathematics, cellular, molecular, *in silico* and *in vitro* molecular interactions, and pathway modeling to understand regulation of molecular networks. Approach facilitates bidirectional flow of knowledge: *in vitro* to *in silico* (Aim 1), *in silico* to *in vitro* (Aim 2), and cell models enabling simulation of complex network dynamics (Aim 3). The approach builds on the collective expertise in immunology, genetics, and applied and theoretical statistics.

Statement of Plans for the Upcoming Research Period

Goal 1. Obtain data from public sources for use in simulations and data model building. Milestone 1A. Access datasets.

Goal 2. Initiate mathematical model building using publicly available datasets. Milestone 2A. Build models to describe the data. Milestone 2B. Test and refine data models.

Literature Cited:

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SC, Jenkins SC, Palmer SM, et al. (1994) A genome-wide search for human type 1 diabetes susceptibility genes. Nature 371:130-136.

Devlin B, Roeder K. (1999) Genomic control for association studies. Biometrics 55:997-1004.

Devlin B, Roeder K, Wasserman L. (2001) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60:155-166.

Devlin B, Bacanu SA, Roeder K. (2004) Genomic Control to the extreme. Nat Genet 36:1129-1130.

Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA. (2005) Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet 76:592-608.

Lander ES, Schork NJ. (1994) Genetic dissection of complex traits. Science 265:2037-2048.

Lee AB, Luca D, Klei L, Devlin B, Roeder K. (2010) Discovering genetic ancestry using spectral graph theory. Genet Epidemiol 34:51-59.

Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M. (2008) On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82:453-463.

Nagelkerke NJ, Hoebee B, Teunis P, Kimman TG. (2004) Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. Eur J Hum Genet 12:964-970.

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Stirnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, Bustamante CD, Roses AD, Burns DK, Ehm MG, Lai EH. (2008) The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83:347-358.

In our third quarterly scientific progress report (03/01/10 – 05/31/10) we then reported the following findings:

Our efforts during the recently completed research quarter were focused upon two goals: 1) To obtain data from public sources for use in simulation and data model building, and 2) To initiate mathematical model building using publicly available datasets. These goals have been completed and are described below. Moreover, our manuscript submitted to ANNALS OF STATISTICS has been accepted for publication and is *in press*. The publication describes our ongoing effort to exploit advanced statistical methods for combining data garnered from different study designs, such as, family-based and case-control studies. The goal of this work has been to develop methodology for combining the results of different study designs into a single test statistic. Application of our method to Type 1 Diabetes identified increased association between gene and disease phenotype by combining inheritance of HLA-class II alleles within families with that collected from unrelated case and control subjects. The citation for this publication is Crossett A, Kent BP, Klei L, Ringquist S, Trucco M, Roeder K, Devlin B. Using Ancestry Matching to Combine Family-Based and Unrelated Samples for Genome-Wide Association Studies. ANNALS OF STATISTICS (*in press*).

<u>Previous Quarter Research Goals</u>
Goal 1. Obtain data from public sources for use in simulations and data model building.
Goal 2. Initiate mathematical model building using publicly available datasets.

The two goals have been completed. Briefly, we have identified appropriate gene expression data accessible through the National Institutes of Health (NIH) sponsored Gene Expression Omnibus (GEO) database (Edgar et al., 2002). This resource consists of primary data on the abundance of mRNA obtained during research projects that have been awarded funding through the NIH. A dataset created using human liver cells (deposited in part by our University of Pittsburgh colleague Stephen Strom) has been identified for our analyses. These data have been downloaded (completing Goal 1) and along with phenotype information available on the human subjects recruited for the study are being used to initiate mathematical modeling of gene-gene interactions described in Goal 2 of the recently completed research quarter.

Detailed Description of Goals 1 and 2

Completion of Goal 1. Obtain data from public sources for use in simulations and data model building.

Table 1. Cohort Summary.

Number of Samples	427
Male/Female Ratio	234/193
Mean Age (years)	50
Age Range (years)	0-94
Number with Steatosis	105

Ref: Schadt (2008) PLOS Biology 6:1020-1032.

Milestone 1A. Access datasets. We have chosen to begin our analyses of gene co-expression networks by evaluating various statistical approaches. To accomplish this we have identified a data set available at the NIH sponsored GEO database. The collection of data consists of microarray generated gene expression data collected from liver samples from n=427 human subjects (Table 1). These data are appropriate for our research into Diabetes and Diabetes Complications in that liver represents an important metabolic organ responding to insulin. Under physiologic conditions in which insulin secretion and/or insulin signaling are dysregulated the liver's contribution to the increase in blood glucose as well as cholesterol and triglycerides is increased. Directly implicating insulin dependent regulation of liver metabolic function with cardiovascular complications associated with Diabetes. The cohort available for this study consisted of n=427 subjects of which approximately 55% where male and 45% female. Notably 25% of the liver samples showed evidence of mild to severe steatosis, that is, fatty liver disease. The steatosis phenotype will be used in subsequent analyses to determine whether gene co-expression correlations are sensitive indicators for the presence of the fatty liver disease phenotype.

Table 2. Example of Genes Tested.

Array ID	Gene ID	Symbol
10033668539	341	*APOC1*
10023813203	348	*APOE*
10023833467	1583	*CYP11A1*
10033668886	1584	*CYP11B1*
10023818702	4547	*MTTP*
10025910281	5105	*PCK1*
10033668775	10891	*PPARGC1A*

As an example of the data collected we obtained data for the complete cohort of n=427 subjects measuring mRNA abundance on roughly 40,000 transcripts from each sample. This corresponds to about 17 million (=427x40,000) individual data. Using customized computer scripts we have organized the data to evaluate the interaction between liver specific target genes and their corresponding transcription factors. As shown in Table 2 select genes were examined based upon their known importance to liver metabolic function. For example, *PCK1* encodes the enzyme Phosphoenolpyruvate Carbolykinase 1 (soluble). *PCK1* gene expression is regulated by insulin. It is produced when insulin levels are low (occurring during fasting) and transcription of

this gene is rapidly turned off when insulin levels are high (during feeding). The gene product is a main control point for regulation of gluconeogenesis. The gene product functions by catalyzing the conversion of oxaloacetate to phosphoenolpyruvate. Other genes examined by our studies were chosen based upon similar criteria of their importance to insulin regulated liver metabolism and metabolic disease.

Table 3. Example of Relative mRNA Abundance.

Array ID	GSM242213	GSM242214	GSM242215
10033668539	-0.3934	-0.3791	-0.1857
10023813203	0.2432	0.1975	0.5117
10023833467	-0.1083	-0.143	-0.0699
10033668886	-0.0322	0.0246	-0.0189
10023818702	-0.192	-0.1284	0.0556
10025910281	0.3219	-0.1297	-0.3554
10033668775	0.154	-0.2382	0.1307

In Table 3 we show an example of the mRNA expression data that has been collected. The table summarizes data for the genes listed in preceding Table 2 (see the column labeled Array ID) but also lists the log(2) normalized expression signal from the mRNA microarray. For the data in Table 3, values of zero indicate no change relative to the mean mRNA abundance while positive values indicate increased mRNA expression and negative values represent decreased gene expression. The column headers beginning with GSM refer to each sample that was used. There is a unique header for each of the n=427 liver samples, providing a data matrix of 40,000-by-427 for co-expression analysis. Collection of this data along with the phenotype information summarized in Table 1 represent completion of Milestone 1A of the previous research quarter.

Completion of Goal 2. Initiate mathematical model building using publicly available datasets.

Milestone 2A. Build models to describe the data. In order to build data models to account for the co-expression correlations between genes we chose to focus on liver specific Transcription Factors that have been shown to regulate genes critical for liver metabolic function, such as those listed above in Table 2. Listed in Table 4 are the Transcription Factors that were evaluated. For each gene, the table includes the official gene symbol, gene identifying number, and full name. For example, *PPARGC1A* encodes the co-transcription factor Peroxisome Proliferator-Activated Receptor Gamma, Coactivator 1 Alpha. The gene product has been previously identified as a critical transcriptional coactivator regulating genes involved in energy metabolism. It is known to interact with other Transcription Factors, such as, cAMP Response Element Binding Protein, Forkhead Box O1, and Peroxisome Proliferator-Activated Receptor Gamma. It provides a direct link between external physiological stimuli and the regulation of mitochondrial biogenesis, and is a major factor regulating cellular cholesterol homeostasis and has been implicated in the development of obesity.

Table 4. Transcription Factors that Regulate Liver Metabolic Functions.

Symbol	Gene_ID	Official Full Name
CEBPA	1050	CCAAT/Enhancer Binding Protein (C/EBP), Alpha
CEBPB	1051	CCAAT/Enhancer Binding Protein (C/EBP), Beta
CEBPD	1052	CCAAT/Enhancer Binding Protein (C/EBP), Delta
CREB1	1385	cAMP Responsive Element Binding Protein 1
CRTC2	200186	CREB regulated transcription coactivator 2
FOXA1	3169	Forkhead Box A1
FOXA2	3170	Forkhead Box A2
FOXA3	3171	Forkhead Box A3
FOXO1A	2308	Forkhead Box O1
HNF1A	6927	HNF1 Homeobox A
HNF4A	3172	Hepatocyte Nuclear Factor 4, Alpha
MLXIPL	51085	MLX Interacting Protein-Like
NR1H2	7376	Nuclear Receptor Subfamily 1, Group H, Member 2
NR1H3	10062	Nuclear Receptor Subfamily 1, Group H, Member 3
NR1H4	9971	Nuclear Receptor Subfamily 1, Group H, Member 4

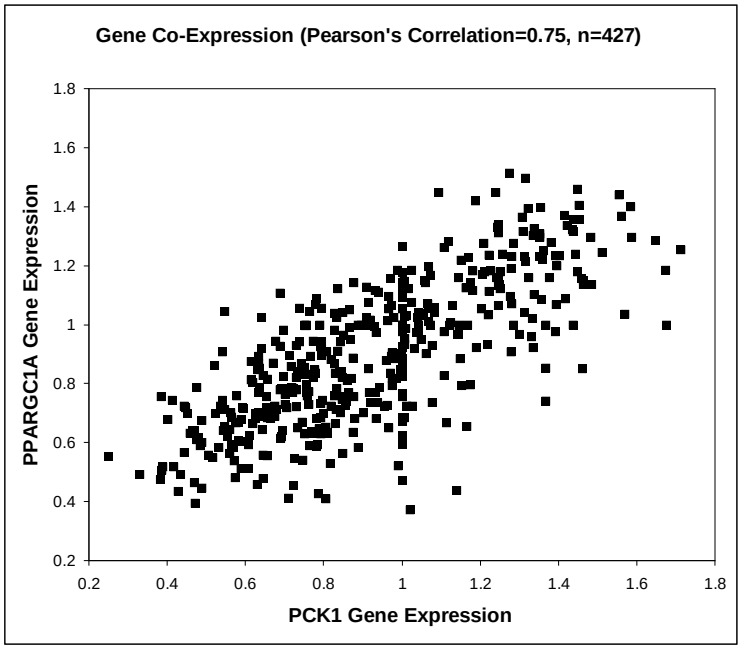
PPARA	5465	Peroxisome Proliferator-Activated Receptor Alpha
PPARD	5467	Peroxisome Proliferator-Activated Receptor Delta
PPARG	5468	Peroxisome Proliferator-Activated Receptor Gamma
PPARGC1A	10891	Peroxisome Proliferator-Activated Receptor Gamma, Coactivator 1 Alpha
RXRA	6256	Retinoid X Receptor, Alpha
RXRG	6258	Retinoid X Receptor, Gamma
SREBF1	6720	Sterol Regulatory Element Binding Transcription Factor 1
SREBF2	6721	Sterol Regulatory Element Binding Transcription Factor 2

Table 5. Selected Transcription Factor and Target Gene Pairs.

Transcription Factor	Gene_ID	Target Gene	Gene_ID
PPARGC1A	10891	*CPT1A*	1374
PPARGC1A	10891	*CYP7A1*	1581
PPARGC1A	10891	*ESRRA*	2101
PPARGC1A	10891	*G6PC*	2538
PPARGC1A	10891	*LDLR*	3949
PPARGC1A	10891	*PCK1*	5105
PPARGC1A	10891	*MEN2*	9927

The Target Genes and Transcription Factors summarized in Table 2 and 4 can also be described in terms of previously identified causal networks. These are summarized in Table 5. For example, the Transcription Factor *PPARGC1A* has been implicated in regulating the expression of Target Genes such as *PCK1* (Phosphoenolpyruvate Carbolykinase 1, soluble) and *G6PC* (Glucose-6-Phosphatase, Catalytic Subunit). The gene product of G6PC is an integral membrane protein of the endoplasmic reticulum that catalyzes hydrolysis of D-glucose 6-phospate to D-glucose and orthophosphate. It is a key enzyme in glucose homeostasis, functioning in gluconeogenesis and glycogenolysis, processes tightly controlled by insulin secretion and signalling.

Milestone 2B. Test and refine data models. The gene expression data collected from n=427 liver samples has been queried for co-expression correlations based upon previously known Transcription Factor and Target Gene interactions summarized in Table 5. As illustrated in Figure 1 we have observed strong correlation between the transcription co-activator *PPARGC1A* and a main enzymatic control point for the regulation of gluconeogenesis, the gene *PCK1*. The Pearson's correlation between these two genes is 0.75 and accounts for roughly 56% of the variance observed in the data (Figure 1). Other gene-gene transcription correlations have been examined. For example, *PPARGC1A* and *G6PC* exhibit a positive correlation (Pearson's correlation measurement equals 0.68) and the insulin responsive transcription factor *FOXO1A* and *PPARGC1A* exhibit a Pearson's correlation of 0.72 (data not shown). The high level of correlation between these gene pairs (i.e., *FOXO1-PPARGC1A*, *PPARGC1A-PCK1*, and *PPARGC1A-G6PC*) is consistent with known insulin regulated transcriptional control of hepatic gluconeogenesis.



Statement of Plans for the Upcoming Research Period

Goal 1. Continue to develop mathematical models to measure gene expression and the correlation between expressions of paired genes.
Milestone 1A. Complete analysis of transcription co-expression observed between transcription factors and their target genes.
Milestone 1B. Discover the presence of new gene-gene co-expression pairs using transcription factors and the complete set of gene expression data.

Goal 2. Initiate the writing of a scientific article for publication in a peer reviewed scientific journal.
Milestone 2A. Outline a manuscript describing the results of our research into recognizing co-expression correlations between genes.

Literature Cited:

Crossett A, Kent BP, Klei L, Ringquist S, Trucco M, Roeder K, Devlin B. Using Ancestry Matching to Combine Family-Based and Unrelated Samples for Genome-Wide Association Studies. ANNALS OF STATISTICS (in press).

Edgar R, Domrachev M, Lash AE. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207-210.

Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusis AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich R. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6:e107.

In the fourth and final quarterly scientific progress report (06/01/10 - 08/26/10) of year 03, we now report on our cumulative results.

Our work performed during the recently completed research quarter focused on the goals of developing mathematical models to measure gene expression and co-expression of correlated genes, and to begin the process of writing the results for publication in a peer reviewed scientific journal. These goals were designed to allow our research group to establish sufficient expertise in the disciplines required to enable the investigation of gene co-expression networks. We have made substantial progress toward achieving both goals as detailed below.

Previous Quarter Research Goals

Goal 1. Continue to develop mathematical models to measure gene expression and the correlation between the expressions of paired genes.

Goal 2. Initiate the writing of a scientific article for publication in a peer reviewed scientific journal.

Detailed Description of Goals 1 and 2

Completion of Goal 1. Continue to develop mathematical models to measure gene expression and the correlation between the expressions of paired genes.

Milestone 1A. Complete analysis of transcription co-expression observed between transcription factors and their target genes. We have identified 3 datasets containing gene expression data collected from biological materials from human subjects. Two of the datasets use B-lymphoblastoid cell lines (BLCL) and the third measured mRNA expression in human liver samples (summarized in Table 1). The combined samples measured n=791 samples collected from human subjects. These data are available to our research effort, have been downloaded, and are currently being evaluated for gene co-expression events. One goal of the project is to use real biological and experimental data to develop advanced mathematical models for recognizing gene-to-gene correlations.

Table 1. Available Genome Wide Expression Data Collected from Human Subjects

Cell Type	Number of Samples	Genome-Wide Expression	Phenotype	Source	Reference
BLCL	294	Yes	None	HapMap	Nayak et al. (2009)
BLCL	70	Yes	Type 1 Diabetes	B.O.Boehm	Personal Communication
Liver	427	Yes	Fatty Liver Disease Type 2 Diabetes	Merck	Schadt et al. (2008); Yang et al. (2010)

The dataset available from human liver samples collected from n=427 subjects is being studied. These data are accompanied by demographic data including subject age, gender, race, and body mass index (BMI) (Table 2). We also have data collected on liver disease, including fatty liver (i.e., steatosis) as well as exposure to liver toxins (data not shown). An early goal that will be pursued in the upcoming research period will be to increase the dimensionality of our analyses by incorporation of subject health and demographic data into our modeling of co-expression networks. The goal of the project is to identify gene neighborhoods that correspond with human health and disease, e.g., steatosis.

Table 2. Demographics of Selected Human Liver Donors

Donor ID	Age	Sex	Race	BMI (kg/m^2)
Steatosis: Mild				
2220015	35	F	W	26.2
2220017	53	M	W	16.3
2220018	22	M	H	22.0
2220021	20	M	W	23.5
Steatosis: Moderate				
2220109	37	M	W	22.4
2220137	61	F	W	33.1
2220138	49	F	W	39.0

2220151	57	M	W	27.7
Steatosis: Severe				
2220074	13	M	W	32.2
2220187	18	M	W	24.9
2220025	28	M	W	26.6
2220191	30	F	W	40.2

Milestone 1B. Discover the presence of new gene-gene co-expression pairs using transcription factors and the complete set of gene expression data. In addition to the samples collected from human liver donors we have comparable datasets on gene expression in BLCL. By combining these data we intend to examine gene networks from n=364 BLCL samples. The samples available from our collaboration with Bernhard Boehm (University of Ulm, Germany) include gene expression data collected from n=3 subjects with Type 1 Diabetes (T1D) and n=4 control subjects. The BLCL were treated with 5 growth conditions (including PMA, LPS-low, LPS-high, and IL-1b) in order to examine stimulation of NFkB transcription factor dependent gene expression during gene network analysis. Table 3 summarizes the experimental design used to stimulate the NFkB dependent gene expression pathway of T1D and control samples. The subject identifying number and phenotype are listed in columns 1 and 2. Columns 3 through 7 list the dataset identifying number for the control growth condition (column 3) as well as stimulating conditions (columns 4 through 7). For each of the 70 conditions that were tested there is a corresponding dataset of roughly 40,000 data points from which the level of a corresponding mRNA transcript can be determined. It is the correlation network existing between these data points that will be the subject of the upcoming work period.

Table 3. Genome Wide Expression Data Collected on T1D and Control BLCLs

Subject ID	Diagnosis	Media	PMA (30ng/ml)	LPS (20ng/ml)	LPS (100ng/ml)	IL-1b (1ng/ml)
Incubation Time 8 Hours						
169B	Control	5299187028H	5299187028J	5299187026B	5299187012B	5299187012I
BOB-5013	Control	5299187028I	5299187028E	5299187030G	5299187030D	5299187029D
ET-2036o	Control	5299187028L	5299187028K	5299187028G	5299187028C	5299187028F
ET-2036w	Control	5299187026F	5299187012C	5299187030E	5299187029H	5299187012D
BOB-5014	T1D	5299187028B	5299187028A	5299187029L	5299187029A	5299187012H
ET-2037o	T1D	5299187029C	5299187026A	5299187026E	5299187030L	5299187012E
ET-2037w	T1D	5299187030A	5299187030B	5299187012A	5299187029I	5299187012G
Incubation Time 24 Hours						
169B	Control	5299187030C	5299187030F	5299187026H	5299187030J	5299187029K
BOB-5013	Control	5299187029J	5299187029E	5299187012L	5299187012F	5299187026J
ET-2036o	Control	5299187027J	5299187027H	5299187027G	5299187027I	5299187029B
ET-2036w	Control	5299187030I	5299187012K	5299187026L	5299187030H	5299187030K
BOB-5014	T1D	5299187026C	5299187026K	5299187012J	5299187026I	5299187026D
ET-2037o	T1D	5299187027C	5299187027E	5299187027L	5299187027F	5299187027A
ET-2037w	T1D	5299187027K	5299187028D	5299187027D	5299187027B	5299187029G

We have been able to identify evidence for as many as n=419 known and predicted NFkB target genes. A subset of known NFkB dependent genes are listed in Table 4. The signals obtained from probes designed to measure the expression of these genes will be used during our initial analyses of the BLCL derived data that were collected and are summarized in the preceding Table 3. The hypothesis being tested is that the stimuli used in the experiment (i.e., the NFkB stimulating compounds PMA, LPS, and IL1b) will show evidence of differentially affecting gene expression of a subset of mRNA transcripts listed in the Table 4 and that the pattern of co-expression will differ between T1D and control subject.

Table 4. Selected NFkB Target Genes

Official Symbol	Gene ID	Official Full Name
Transcription Control		
AHCTF1	25909	AT hook containing transcription factor 1
CEBPD	1052	CCAAT/enhancer binding protein (C/EBP), delta
CREB3	10488	cAMP responsive element binding protein 3
E2F3	1871	E2F transcription factor 3
LEF1	51176	lymphoid enhancer-binding factor 1
SP7	121340	Sp7 transcription factor
STAT5A	6776	signal transducer and activator of transcription 5A
TFEC	22797	transcription factor EC
YY1	7528	YY1 transcription factor
Chemokines and Chemokine Receptors		
CCL5	6352	chemokine (C-C motif) ligand 5
CCR5	1234	chemokine (C-C motif) receptor 5
CCR7	1236	chemokine (C-C motif) receptor 7
Cell Division/Cell Cycle Control		
CCND1	595	cyclin D1
CCND2	894	cyclin D2
Interleukin Cytokines		
IL1B	3553	interleukin 1, beta
IL2	3558	interleukin 2
IL8	3576	interleukin 8

Completion of Goal 2. Initiate the writing of a scientific article for publication in a peer reviewed scientific journal.

Milestone 2A. Outline a manuscript describing the results of our research into recognizing co-expression correlations between genes. The research being carried out as a result of this DOD funding mechanism has been greatly enhanced by our collaborations. In particular, the study of gene co-expression networks is being pursued in close collaboration with Kathryn Roeder's research team at Carnegie Mellon University. As a result of our work together her group has published a paper describing methods they have developed for working with high dimensional datasets (Liu et al., 2010). In that manuscript they present a new method for detecting gene co-expression correlations. The methodology generates with high probability true interactions and is independent of sample size and dimensionality of the dataset (Figure 1). Prior to development of the current method, standard techniques used Bayesian information criteria (BIC). Unfortunately, BIC can perform poorly when dimensionality of a dataset is large relative to sample size as commonly occurs when working with gene expression data.

In contrast, the new approach (denoted by the acronym StARS) chooses the network regularization parameter so that the resulting network is sparse without excessive variability across subsets of the network. This is accomplished by incrementally reducing the level of regularization while monitoring variability between sub-samples. Regularization is performed until the variability between the resulting networks is minimized. In applying the method to problems like gene regulatory networks the aim is to investigate the interaction of many genes (see Figure 1 for an illustration). We have chosen to tolerate a few false positive interactions so long as false negatives are minimal (notice that StARS versus BIC results in many fewer false edges). In contrast, to our method the BIC approach frequently recognizes a very large number of false positives. Thus, creating a difficult situation when designing biological experiments to test the conclusions predicted by the network model.

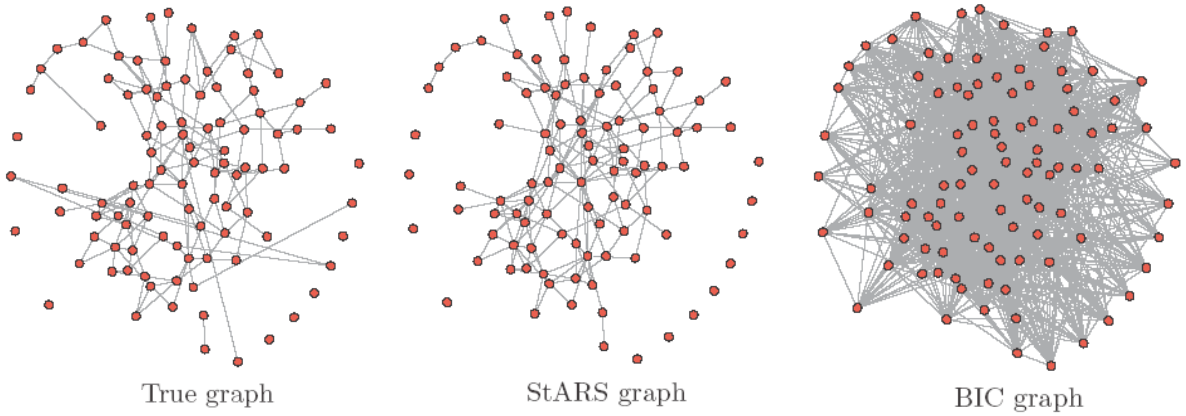


Figure 1. Comparison of the StARS method developed during the DOD funded project with BIC. Data used assumed a sample size of n=400 and dimensionality of the data equal to 100. Note the high degree of similarity between edges defined by the True graph and the StARS approach compared with the excessive number of edges identified using the BIC method.

Statement of Plans for the Upcoming 6-Month Research Period

Goal 1. Using the data available on human BLCL and liver samples continue to develop mathematical models for identifying gene co-expression networks.

Milestone 1A. Analyze gene expression data collected from human samples for the presence of gene networks.

Milestone 1B. Incorporate human phenotype and demographic data into the analysis of co-expression networks.

Goal 2. Incorporate gene expression and phenotype data from animal models into the analysis of gene co-expression networks.

Milestone 2A. Identify gene expression datasets collected from mouse models of human disease, including diabetes.

Literature Cited:

Liu H, Roeder K, Wasserman L. (2010) Stability approach to regularization selection (StARS) for high dimensional graphical models. http://arxiv4.library.cornell.edu/pdf/1006.3316

Nayak RR, Kearns M, Spielman RS, Cheung VG. (2009) Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. Genome Res 19:1953-1962.

Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusis AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich R. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6:e107.

Yang X, Zhang B, Molony C, Chudin E, Hao K, Zhu J, Gaedigk A, Suver C, Zhong H, Leeder JS, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich RG, Slatter JG, Schadt EE, Kasarskis A, Lum PY. (2010) Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. Genome Res 20:1020-1036.

KEY RESEARCH ACCOMPLISHMENTS:

1. Creation of a data repository containing gene expression data from greater than n=700 human subjects and 40,000 genes for each subject.

2. Development of mathematical models for recognizing and evaluating the correlations between co-expressed genes.

3. Identification of gene networks and network neighborhoods correlating with human disease phenotypes.

4. Publication of 6 manuscripts.

REPORTABLE OUTCOMES:

Manuscripts (6 publications)

1. Lu, L., Boehm, J., Nichol, L., Trucco, M., and Ringquist, S. Multiplex HLA typing by pyrosequencing. In Methods in Molecular Biology, vol 496: DNA and RNA Profiling in Human Blood. ed. P. Bugert. Humana Press Inc., Totowa, New Jersey (2009).
2. Kim, D.H., Ringquist, S., and Dong, H.H. Fructose - A sweet risk of fatty liver disease. In: Chocolate, Fast Foods and Sweeteners: Consumption and Health. ed. M.R. Bishop. Nova Publishers Inc., (2010).
3. Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. Screen and clean: a tool for identifying interactions in genome-wide association studies. Genetic Epidemiology 34, 275-285 (2010).
4. Kim, D.H., Zhang, T., Ringquist, S., and Dong, H.H. Targeting FoxO1 for hypertriglyceridemia. Current Drug Targets (in press).
5. Kamagate, A., Kim, D.H., Zhang, T., Slusher S., Strom, S.C., Bertera, S., Ringquist, S., and Dong, H.H. FoxO1 links hepatic insulin action to endoplasmic reticulum stress. Endocrinology 151, 3521-3535 (2010).
6. Crossett, A., Kent, B.P., Klei, L., Ringquist, S., Trucco, M., Roeder, K., and Devlin, B. Using ancestry matching to combine family-based and unrelated samples for genome-wide association studies. Annals of Statistics (in press).

Development of Cell Lines, Tissue or Serum Repositories
1. Repository of DNA samples collected from T1D and T1DN patients exceeding 1,800 subjects.

CONCLUSION:

The conclusions from the current year of funding are that mathematical models designed to recognize gene co-expression correlations as well as gene networks and highly interconnected sub-networks can be accomplished with high probability of identifying true correlations. Correlated gene pairs and neighborhood groups will, in turn, provide evidence for their role in human disease phenotypes. The work planned in the upcoming year will test the hypothesis that study of co-expression network in cells isolated from diabetic patients when compared with cell collected from healthy subjects will identify disease specific group of genes and reveal molecular genetic pathways leading to disease susceptibility.

The research project generated 6 publications. These are listed under the section entitled "REPORTABLE OUTCOMES".

The So What Section. What are the implications of this research? Diabetes affects 16 million Americans and 800,000 new cases annually. African, Hispanic, Native and Asian Americans are particularly susceptible to its most severe complications. Costs associated with diabetes may be as high as $132 billion. Diabetes accounts for 42% of new cases of end-stage renal disease with over new 100,000 cases per year at an average cost of $55,000 per patient annually.

What are the military significance and public purpose of this research? As the military is a reflection of the U.S. population improved prediction of risk for developing diabetes and diabetic complications among active duty

members of the military, their families, and retired military personnel will potentially allow focused preventative treatment of at risk individuals, providing significant healthcare savings and improved patient well being.

REFERENCES:

1. Lu, L., Boehm, J., Nichol, L., Trucco, M., and Ringquist, S. Multiplex HLA typing by pyrosequencing. In Methods in Molecular Biology, vol 496: DNA and RNA Profiling in Human Blood. ed. P. Bugert. Humana Press Inc., Totowa, New Jersey (2009).

2. Kim, D.H., Ringquist, S., and Dong, H.H. Fructose - A sweet risk of fatty liver disease. In: Chocolate, Fast Foods and Sweeteners: Consumption and Health. ed. M.R. Bishop. Nova Publishers Inc., (2010).

3. Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. Screen and clean: a tool for identifying interactions in genome-wide association studies. Genetic Epidemiology 34, 275-285 (2010).

4. Kim, D.H., Zhang, T., Ringquist, S., and Dong, H.H. Targeting FoxO1 for hypertriglyceridemia. Current Drug Targets (in press).

5. Kamagate, A., Kim, D.H., Zhang, T., Slusher S., Strom, S.C., Bertera, S., Ringquist, S., and Dong, H.H. FoxO1 links hepatic insulin action to endoplasmic reticulum stress. Endocrinology 151, 3521-3535 (2010).

6. Crossett, A., Kent, B.P., Klei, L., Ringquist, S., Trucco, M., Roeder, K., and Devlin, B. Using ancestry matching to combine family-based and unrelated samples for genome-wide association studies. Annals of Statistics (in press).

APPENDICES:
None